Capstone Project Report: accident severity

1) Introduction/Business Problem

The health service in Seattle City want to understand the amount of emergency vehicles and staff they should hire to deal with road traffic accidents. The health service in Seattle City want to know if a model can be used to predict the severity of road traffic accidents. They want to use this model to optimise the number of support vehicles and staff they have to deal with the accidents. This can help the health services of Seattle City optimise their resources and have a better chance of saving lives.

I have researched and found a data set which looks at the volume and severity of road traffic accidents in Seattle City between 2004 and 2020. Using Machine Learning techniques I will build a series of classification models in order to predict the severity of an accident. I will test a series of logistic regression and decision trees models in order to understand which model predicts the severity of an accident most clearly. Understanding the factors which predict the severity of accidents can help the health services of Seattle City plan for accidents more effectively.

Target Audience: The health and emergency services of Seattle City

2) Data Section

I have researched and found a data set which looks at the volume and severity of road traffic accidents in Seattle City between 2004 and 2020. The data set is comprised of over 194,000 traffic accidents in Seattle City in that time period and for each accident we have whether the accident was property damage only or whether it also included an injury to a person. There is lots of other information supplied also, this includes the features such as the date of the traffic accident, the location of the accident, the collision type, the number of people involved, the vehicle type, the weather at the time of the accident, if the accident took place in daylight and more. All of these features will be tested to understand if they help to predict the severity of an accident.

Using this information I plan to test a series of machine learning models in order to predict traffic accident severity in Chicago city. I will use a number of machine leaning classification techniques include logistic regression, decision trees and support vector machines in order to understand which model predicts accident severity with the most accuracy. Once the models are built, we will be able to understand which of the features above are most important to predict the severity of road traffic accidents in Seattle City. This will help the health service better plan their resources, for example if they know that injuries are more likely to happen in wet weather they can have more support staff and vehicles ready in wetter months.

Seattle City data source:

Seattle city accident data: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

A descriptive summary of the data source can be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

3) **Methodology**

For this project I am using a Python programming software and libraries to investigate the data and build models to help predict accident severity.

The dataset required cleaning as there were missing rows and some of the columns were dropped as they gave no additional information (ie report number provides no explanation). A number of the variables were also in a categorical format, these were transformed into binary code as this is what classification models require.

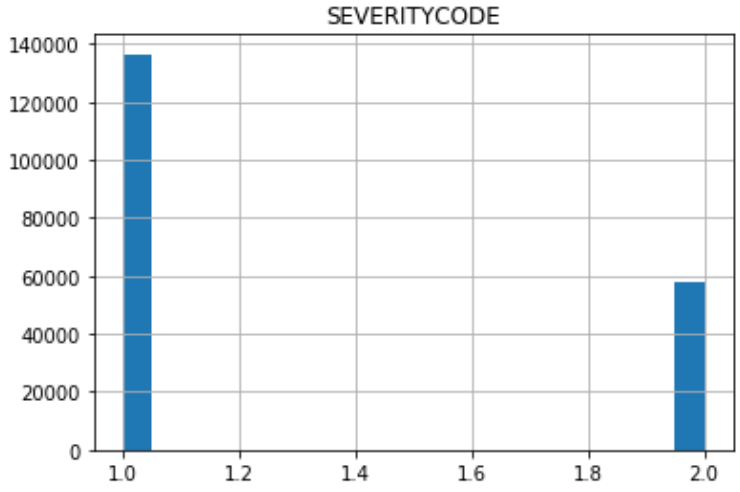
In this problem we are trying to predict if the severity of an accident is either 1 – property damage or 2 – an injury to a person. This is a classification problem and I decided to test two Machine Learning techniques to predict the severity of the accident.

Before building the models it is necessary to explore the data. Here is a histogram which shows the total number of accidents. There were around 136k property damage accidents (1) and close to 60k accidents involving personal injury (2) recorded in the dataset.

```
Name: SEVERITYCODE, dtype: int64
```

```
[24]: df.hist(column='SEVERITYCODE', bins=20)
```

```
[24]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000021B54236E48>]],
      dtype=object)
```




```
[11]: #Clean data
```

After this I also did a correlation analysis to understand which of the variables are correlated with the two types of accident severity. I then chose a number of variables to use to build models to predict accident severity. The variables are:

Weather

Daylight or Night-time

If the vehicle was speeding at the time of the accident

If a bicycle was involved in the accident

If the road was wet, icy or dry

If there was fog when the accident happened

The type of accident and the number and vehicles people involved

Models

I used two classification techniques for this project: decisions trees and logistic regression. Both techniques are popular machine learning algorithms which build a model from the historical data of accidents and then predict the severity class of an unknown accident given information. These techniques are required in this project as we are using a nominal variable (i.e. predicting whether an accident would either be property damage or personal injury). Linear regression cannot be used as we are not predicting a continuous variable.

My plan is to test the explanatory variables listed above into these frameworks and then chose the framework which provides the best level of accuracy.

The data is then split into a test set and a training set. The model is trained on the training set and then we test the model on the test set to understand the accuracy of the model.

4) **Results**

One method for understanding the accuracy of a model is to use the Jaccard index. This index compares the two sets of data to see which members are shared and which are distinct. It is a measure of the similarity for the two data sets with a range from 0% to 100%. The higher the percentage, the more similar the two data sets.

Decision Tree Jaccard Test: 74.4%

```
Name: SEVERITYCODE, dtype: int64

30]: from sklearn import metrics
     import matplotlib.pyplot as plt
     print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_testset, predTree))

     DecisionTrees's Accuracy:  0.7445635423444402
```

Logistic Regression Jaccard Test: 73.5%

```
88]: from sklearn.metrics import jaccard_similarity_score
     jaccard_similarity_score(y_test, yhat)

     C:\Users\Paul.Young\Anaconda3\lib\site-packages\sklearn\met
     ore has been deprecated and replaced with jaccard_score. It
     ing behavior for binary and multiclass classification tasks
       FutureWarning)

88]: 0.7354308462822653
```

The test results are very similar and both models predict the severity of an accident correctly in the test set around 74% of the time which is a solid result.

I also looked at some other statistical tests and decided that the decision tree model is superior based on these – the decision tree had a higher F1 score (another measure of accuracy)

Decision Tree: 0.71 (d = 6)

	Depth	F1-score	Jacard
	d=3	0.666725	0.732852
	d=4	0.668412	0.740112
	d=5	0.669675	0.740728
	d=6	0.710536	0.744564

Log Regression: 0.68

```
# evaluate Logistic Reg
fs = round(f1_score(y_test, lr_yhat, average='weighted'),2 )
```

```
fs
```

```
0.68
```

Result: the decision tree has greater accuracy and is a better model to predict accident severity.

5) **Discussion**

The solid accuracy score of the decision tree model (74.5%) means that we can use the model to help the health service of Seattle City predict accident severity more effectively.

Using the decision tree we can see that accident severity increases in daylight, when cars are speeding and road conditions are wet or icy. The health service of Seattle City can use this information to have more vehicles and staff ready in months where rainfall and ice is more likely. They can also lobby the local government to increase the severity of speeding fines and introduce new signs on roads to tell people to slow down.

In order to update and improve the model we can also work with the health service in Seattle to track other data which may help improve the accuracy of the model. This could include the type and age of the vehicles involved, if the vehicle has airbags and if passengers were wearing seatbelts at the time of the accident. Furthermore, we could look more deeply into accident severity to assess the type of injury and if certain equipment was needed at the scene of the accident. This information could improve the accuracy and results of the analysis further and give more information to the Seattle City health service.

6) **Conclusion**

Using accident severity data we have built a robust classification model to help the health service of Seattle predict the severity of an accident given certain conditions such as wet and icy weather. This information can help the health service plan their resources and respond more effectively to accidents. Modelling is an iterative process and there may be other information which can help to improve the accuracy of the model further.